Welfare Maximization in Competitive Equilibrium: Reinforcement Learning for Markov Exchange Economy

Zhihan Liu[1], Miao Lu[2], Zhaoran Wang[1], Michael Jordan [3], and Zhuoran Yang[4]

[1]Northwestern University [2]University of Science and Technology of China,[3]UC Berkeley, [4]Yale University

Overview

In this paper, we propose a novel bilevel economic model Markov Exchange Economy (MEE) and provide two provably efficient algorithms to solve the online version and offline version of MEE, respectively.

Markov Exchange Economy (MEE)

- **Exchange Economy (EE):** A set of rational agents with individual initial endowments allocate and exchange a finite set of valuable resources based on a common price system. The target of EE is to achieve Competitive Equilibrium (CE), where agents maximize their own utilities *under their budget constraint*.
- **Social Welfare Maximization (SWM):** A *central planner* is introduced in EE to steer the system so as to achieve *Social Welfare Maximization (SWM)* that can be defined as the sum of the utilities of all agents over the entire episode.
- **Dynamic Model:** A finite horizon MEE consists of N agents, one social planner, and H time steps. Each state s_h consists a context c_h and endowments e_h. The transition of the next state is only determined by the current state and the action of the planner.
- **Joint Optimality:** The joint optimal policy of MEE satisfies both CE and SWM.

MOLM for Online MEE

- **Model estimation step:** We construct confidence sets $\mathcal{U}_h^k$ for utility functions and $\mathcal{P}_h^k$ for transition kernels using data from previous $k-1$ episodes.
- We use value targeted regression (VTR, Ayoub et al., 2020) for transition estimation.
- **Optimistic planning step:** We use $\mathcal{U}_h^k$ and $\mathcal{P}_h^k$ to perform optimistic planning to approximate the joint optimal policy:

$$\nu_h^k(s) = \texttt{CE}(\{\hat{u}_h^{k,(i)}(s,\cdot)\}_{i\in[N]}),$$

$$\pi_h^k(s) = \arg\max_{b\in\mathcal{B}} \sum_{i=1}^{N} \int_{\mathcal{S}} V_{h+1}^{k,(i)}(s')\widehat{P}_h^k(\mathrm{d}s'|s,b),$$

where $\hat{u}_h^k \in \mathcal{U}_h^k$ and $\widehat{P}_h^k \in \mathcal{P}_h^k$ are optimistic estimations.

Analysis of MOLM

- **Sublinear Regret.** Online regret of MOLM for K episodes:

$$\mathrm{Regret}_{\mathrm{CE,SWM}}(K) \in \tilde{\mathcal{O}}(H^2N\sqrt{dK}),$$

where d is the eluder dimension of the function classes for general function approximations (Russo & Van Roy, 2013).

MPLM for Offline MEE

- **Model estimation step:** construct confidence sets $\mathcal{U}_h$ for utility functions and $\mathcal{P}_h$ for transition kernels using previously collected offline data only.
- **Pessimistic policy optimization step:** use $\mathcal{U}_h$ and $\mathcal{P}_h$ to perform pessimistic policy optimization to approximate the joint optimal policy:

$$\hat{\nu}_h(s) = \texttt{CE}(\{\hat{u}_h^{(i)}(s,\cdot)\}_{i\in[N]}),$$

$$(\widehat{\pi},\widehat{P}) = \arg\max_{\pi\in\Pi} \min_{\hat{P}:\{\hat{P}_h\in\mathcal{P}_{h,\xi_2},\forall h\in[H]\}} \sum_{i=1}^{N} \widehat{V}_{1,(\hat{P},\hat{u})}^{(\pi,\hat{\nu}),(i)}(s_1),$$

where $\hat{u}_h \in \mathcal{U}_h$ and $\widehat{P}_h \in \mathcal{P}_h$ are pessimistic estimations.

Analysis of MPLM

- **Global Convergence.** Offline suboptimality of MPLM algorithm:

$$\mathrm{SubOpt}(\widehat{\pi},\hat{\nu}) \in \tilde{\mathcal{O}}(H^2N\sqrt{C^\star\iota/K}).$$

where ι is the covering number of the function classes for general function approximations and $C^\star$ is the concentrability coefficient between the offline dataset $\mathbb{D}$ and joint optimal policy $(\pi^\star,\nu^\star)$. Hence we prove that MPLM efficiently finds the jointly optimal policy $(\pi^\star,\nu^\star)$ approximately.